Science and Technology Infusion Climate Bulletin
NOAA's National Weather Service
38th NOAA Annual Climate Diagnostics and Prediction Workshop
College Park, MD, 21-24 October 2013

Developing a More Reliable and Usable ENSO Prediction Plume

Anthony G. Barnston and Michael K. Tippett
International Research Institute for Climate and Society,
The Earth Institute at Columbia University, Lamont Campus, Palisades, NY

Huug van den Dool and David A. Unger
Climate Prediction Center, NCEP/NWS/NOAA, College Park, MD

1. Introduction

Since early 2002, the International Research Institute for Climate and Society (IRI) has issued, each month, a collection of the forecasts from a large number of ENSO forecasting institutions, in the form of an ENSO prediction plume (Fig. 1). The forecasts predict the Nino3.4 index in the tropical Pacific (SST averaged over 5ºN-5ºS, 120º-170ºW).

In late 2011 this forecast plume became a product of both IRI and the NOAA Climate Prediction Center (CPC). Although the product has been popular and frequently viewed on the Web, it has had several significant problems:

- The forecast producers do not form their anomalies with respect to the same 30-year base periods as encouraged, and IRI/CPC does not correct for such (usually minor) deviations.
- The forecast spread within individual models, indicative of model uncertainty, is ignored and only the mean forecast is shown.
- Model biases, evident upon examination of hindcasts, are not corrected; and some forecasts are from models that lack hindcasts.
- No attempt is made to provide a final forecast probability distribution; users see the spread of the model forecasts and are left to surmise the uncertainty on their own.

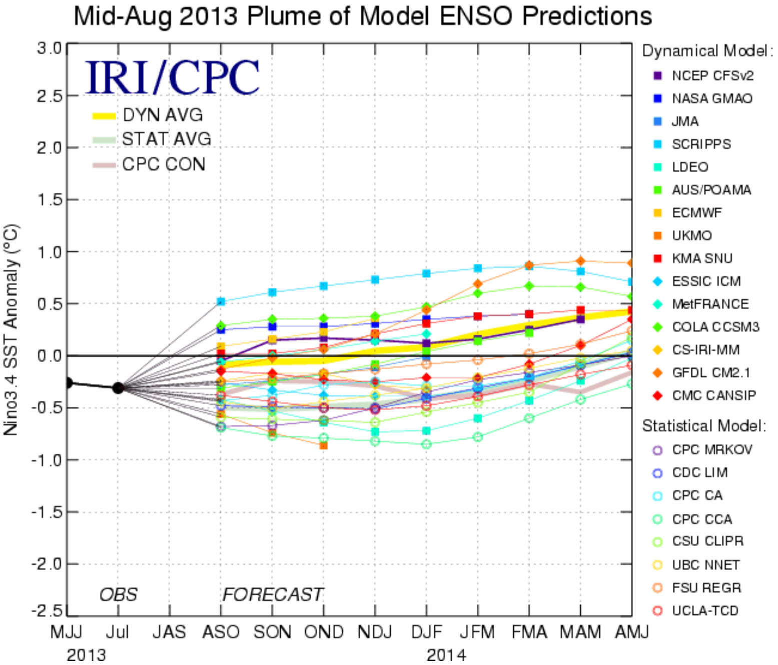

Fig. 1 Example of an IRI/CPC ENSO prediction plume product, issued in mid-August 2013.

Of the four problems listed above, the third one appears most serious, because some of the dynamical models are known to have substantial (>0.5ºC) biases. Hence, some of the spread in the model forecasts shown in Fig. 1, even at very short lead times, may well be due to differing model biases. The ENSO forecast plumes posted on the CPC Web site from the North American Multi-model Ensemble (NMME) project (Kirtman *et al.* 2014) have undergone hindcast-based bias correction by start month and lead time, and the resulting plume is noticeably less wide than the IRI/CPC plume at short leads. The NMME plume also shows all ensemble members of all models, forming a very dense cluster of lines on the plot.

Correspondence to: Anthony G. Barnston, International Research Institute for Climate and Society, The Earth Institute at Columbia University, Lamont Campus, Palisades, NY 10964; E-mail: tonyb@iri.columbia.edu.

The current work attempts to develop a protocol for selecting and processing the incoming forecasts for the IRI/CPC plume so as to eliminate or greatly reduce each of the problems identified above. Because the most serious problem (lack of bias correction) requires a multidecadal hindcast history to evaluate bias, it appears that forecasts from models lacking an adequate hindcast history will not qualify for a higher quality version of the plume.

This work uses as test cases a set of 6 models from the NMME project, because those models all have 29-year hindcast data that is conveniently available. The 6 models include (1) NCAR/Univ. Miami CCSM3 (6 members), (2) NOAA/NCEP CFSv2 (24 members), (3) Canada CMC#1 (10 members), (4) Canada CMC#2 (10 members), (5) NOAA GFDL model (10 members), and (6) NASA model (11 members). All 6 models have a 1982-2010 hindcast period. Models' maximum lead times vary from 9 to 12 months. Besides looking at the forecast characteristics of each model, those of the combined forecast (our MME) are studied. The MME is formed by combining the individual ensemble members of all of the models. Because some models have many more members than others, the number of members acts as an effective weighting system: *e.g.*, the NOAA/NCEP CFSv2 has 4 times as many ensemble members as the NCAR /Univ. Miami CCSM3, so it will exert 4 times the weight of CCSM3 in forming the MME forecast. Here, we forecast 1-month mean SST rather than seasonal mean SST as done in the IRI/CPC plume.

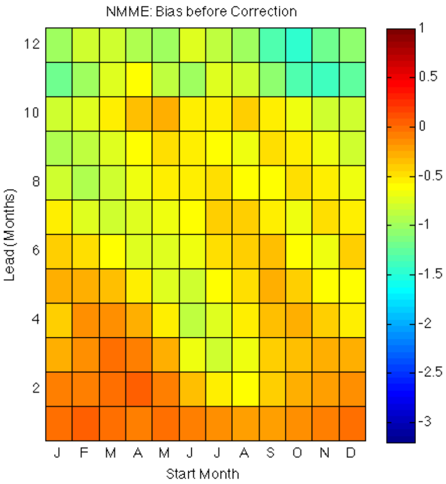

Fig. 2 Bias of the MME in forecasts of Nino3.4 SST, by start month (from Jan to Dec along x-axis) and lead time (from 1 to 12 from bottom to top along y-axis).

2. Results

The basic discrimination skill of each of the 6 models is examined using the temporal correlation (or "anomaly correlation") between Nino3.4 SST hindcast and observation for each start month and each lead time up to 12 months lead. Although the model skill profiles differ from one another in their details, all are seen to have acceptable profiles with the expected seasonal distribution (not shown). However, an examination of mean bias indicates major differences in bias among the models, both in general severity and in distribution over start months and leads. It is clear that each model should be bias-corrected prior to being shown on an improved ENSO prediction plume. The net bias of the MME, shown in Fig. 2, lacks the severity of the biases of individual models due to some bias cancellation, but still reveals a moderate negative bias at long leads and at intermediate leads for some times of the year.

Another kind of bias that individual models may carry is forecast amplitude bias. The interannual standard deviation of the forecasts should not be larger than that warranted by the model's correlation skill, which would be approximately that of the observations multiplied by the skill (Hayes 1973). Such a prescription for the amplitude of the ensemble mean forecast would minimize mean squared errors and produce probabilistically reliable forecasts. However, each model has its "own world", with signal-to-noise ratios that may not agree with that of the real world. It turns out that the amplitude of the MME forecast does not deviate greatly from the ideal amplitude, so that correction of the amplitude by start month and lead does not greatly change the performance of the forecasts. Figure 3 shows the root-mean-square error (RMSE) skill score, defined as 1 – ($RMSE_{fct}$ / $RMSE_{cli}$) where fct refers to the forecasts and cli refers to perpetual climatology forecasts (*i.e.*, zero anomaly). Figure 3 shows that the RMSE is generally substantially improved with bias correction, and only slightly more by forecast amplitude correction.

Although amplitude correction does not change the RMSE skill score dramatically, the amplitude corrections are not minor. Figure 4 shows the MME forecast-to-observation standard deviation ratio before and after correction for the amplitude, and it is clear that the forecasts tend to have too high a standard deviation before correction, especially at intermediate and long leads. While the standard deviation of individual ensemble members is expected to be comparable to that of the observations for all start times at all

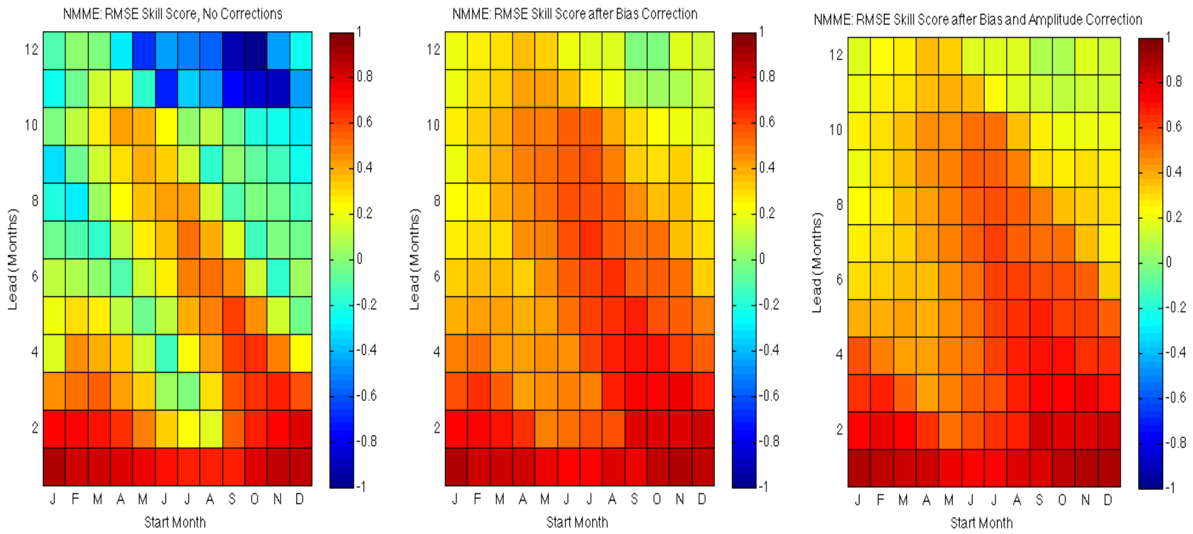

Fig. 3 RMSE skill score for the MME forecasts, by start month (x-axis) and lead time (y-axis). See the text for the definition of the score. The left panel shows skills without any corrections, the middle panel with individual model bias corrections, and right panel with both bias and amplitude corrections.

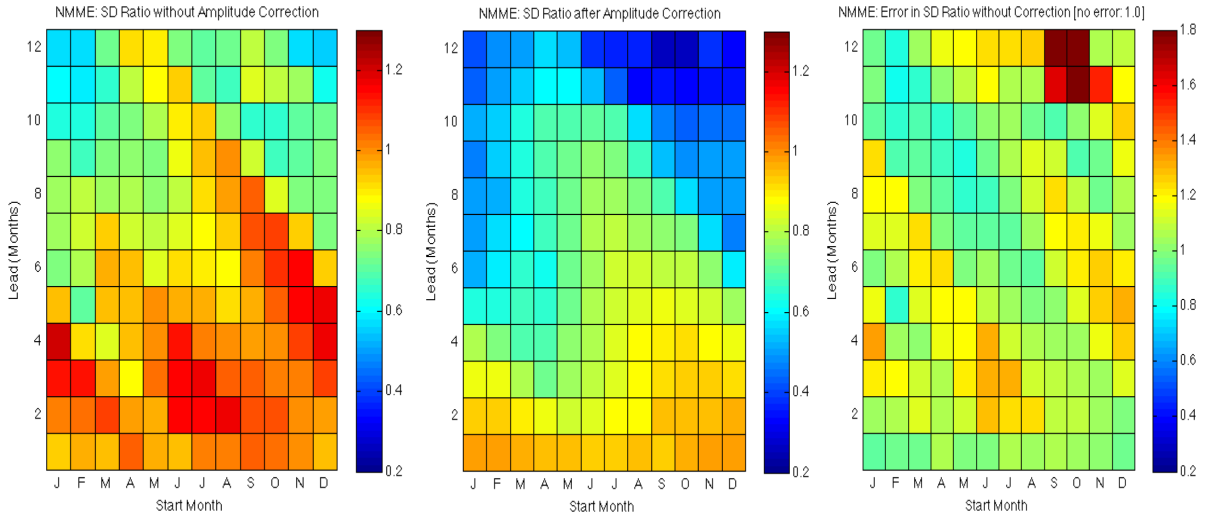

Fig. 4 Standard deviation ratio of MME forecasts versus observations. Left panel shows ratios without amplitude correction, middle panel with amplitude correction, and right panel the ratio of the values without correction to those with correction (note the different scale for the right panel).

leads, that of the ensemble means should be in proportion to the lack of predictability. Although predictability within each model's world is estimated by its signal-to-noise ratio (interannual standard deviation of ensemble mean forecast versus ensemble spread), the actual predictability is better estimated by the correlation between the forecasts and observations. It is this latter measure of realized predictive skill that should govern the interannual standard deviation of the forecasts.

The most appropriate interannual standard deviation for each start month and lead time is determined by the actual temporal correlation skill of the hindcasts with observations, such that a correlation of 0.5 would imply an ideal MME forecast standard deviation of 0.5 that of the observations. Using this indirect way to set the forecast amplitude corrects for model signal-to-noise ratios that do not properly reproduce that in nature.

An important characteristic of a forecast is its uncertainty. For individual forecasts, uncertainty is ideally expressed by the spread of the ensemble members. Even the shape of the distribution of the member forecasts may occasionally be meaningful if it is based on the physics at play in the forecast rather than just accidental

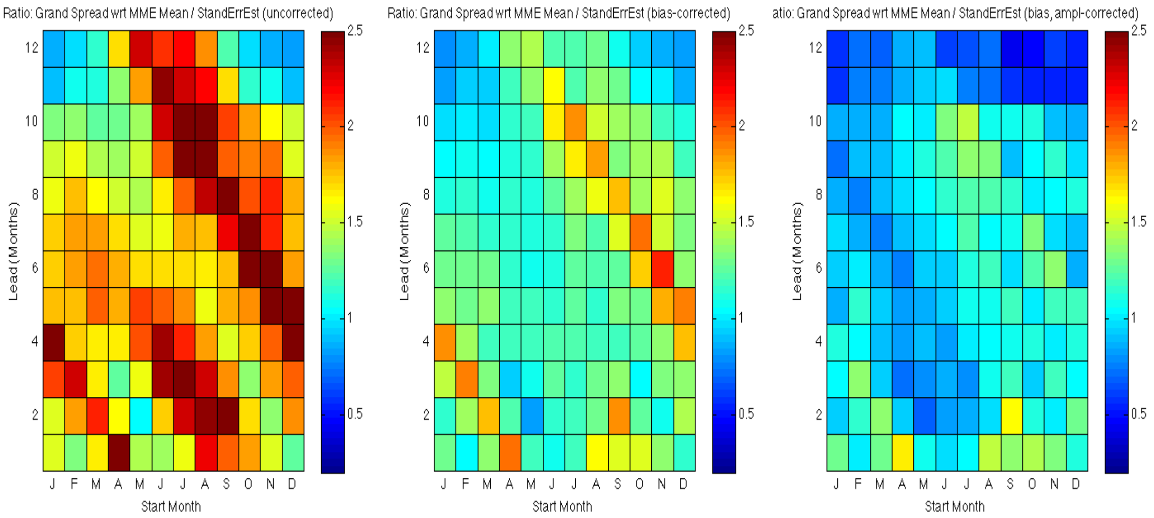

Fig. 5 Ratio of the MME spread (across all model members) to the actual hindcast skill-based standard error of estimate. Left panel shows the ratio for no corrections, middle panel for only bias corrections, and right panel for both bias and amplitude corrections. The ideal ratio is 1.

sampling variability in the finite set of ensemble members. A check for a reasonable magnitude of ensemble spread is the standard error of estimate, based again on the actual correlation skill of the MME forecasts over the hindcast period, by start month and lead time. The standard error of estimate (SEE) is defined by

$$SEE = SD_y \sqrt{1 - cor_{xy}^2}$$

The above formula implies that high-skill forecasts should have a smaller spread than lower-skill forecasts. Is this formula followed to first order in the MME hindcasts? Figure 5 shows the ratio of the MME spread (across all model members) and the skill-based SEE for the cases of no corrections, only bias corrections, and bias and amplitude corrections. The ratios in Fig. 5 indicate far too much spread in model members without bias correction, and a much more realistic spread after bias correction. Further improvement of the ratio (toward 1) occurs with amplitude correction.

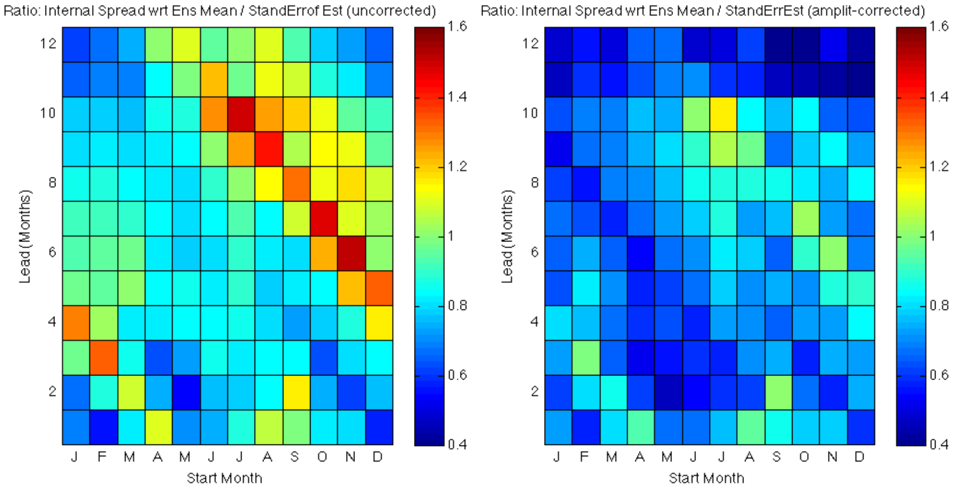

Fig. 6 Spread of the MME forecasts coming from the variation of the members of each model with respect to its own ensemble mean ("internal" model spread). Left panel shows internal spread before amplitude correction, right panel following amplitude correction.

One contribution to the spread of the MME forecasts is that among the members of each model with respect to its own ensemble mean, while a second contribution is that of the differing ensemble means across the models. We ask how much the first component of the spread is contributing to the total spread. Figure 6 shows this aggregated "internal" member spread before and after amplitude correction. The internal spread after the amplitude correction is generally well below the level that is compatible with the SEE.

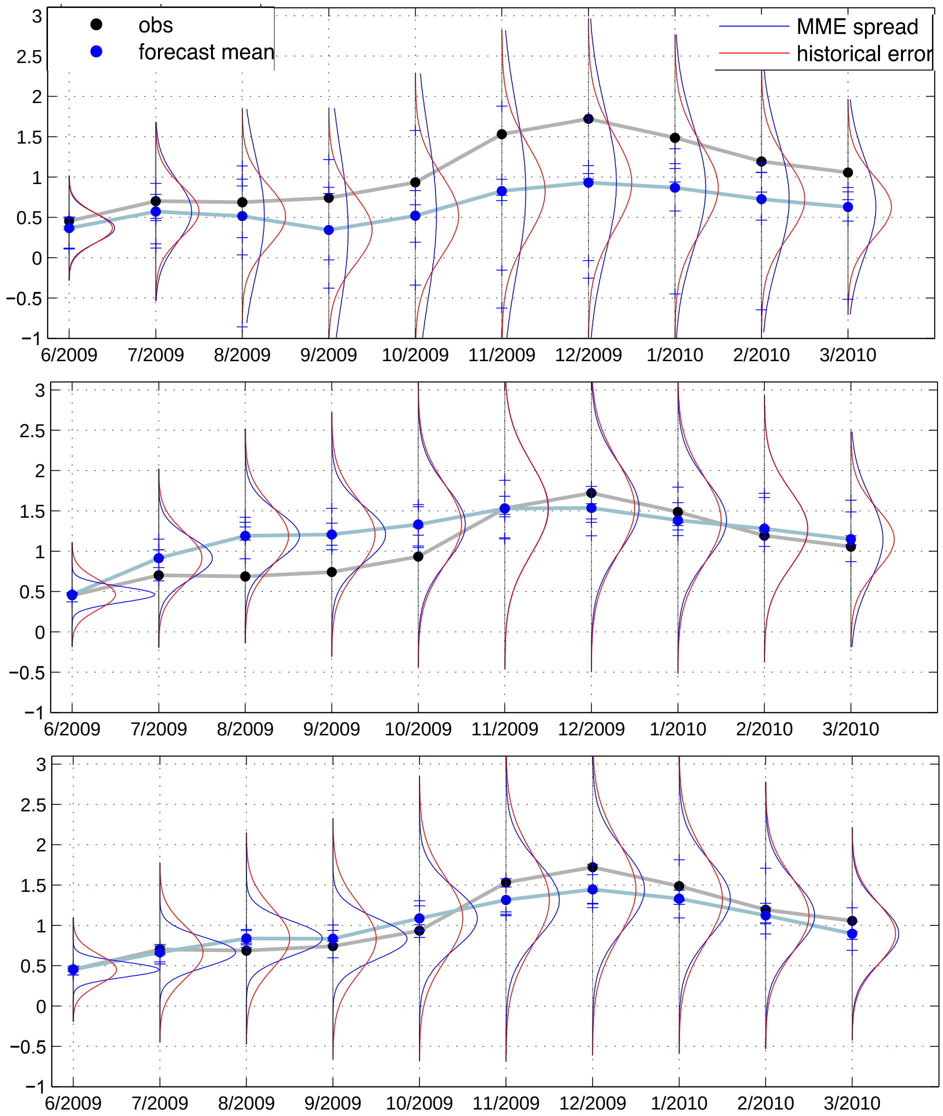

Fig. 7 MME forecasts from June 2009 for the period of the 2009/2010 El Niño event. Top panel shows forecasts without any corrections, middle panel after bias correction, and bottom panel after bias and amplitude correction. The blue line and solid dots show the MME mean forecasts; the black line and dots show the observations. The horizontal ticks on the vertical line for each month show individual model ensemble mean forecasts. The thin blue vertical vertical Gaussian distribution curves show forecast uncertainty based on the MME spread, and the thin red vertical distribution curves show uncertainty based on the hindcast skill-based standard error of estimate.

Figure 7 illustrates an example of the effects of the bias and amplitude corrections in an individual forecast case—here, for forecasts from June 2009 for what turned out to be a moderate strength El Niño during late 2009 and early 2010. Without any correction, the MME forecast substantially underestimates the strength of the event, and the uncertainty is overestimated (especially at short lead times) due to the differing biases of the ensemble means of the various models. Note that the ensemble mean forecasts of the different models differ greatly without any correction. Correction of the mean biases leads to a much improved MME forecast (middle panel) and more realistic width of the uncertainty distribution. Correction for the amplitude as well as the bias results in slight underestimation of the strength of the event, and some underestimation of the amount of uncertainty at short leads. The strength underestimation may be partly a result of the more conservative forecast amplitude following amplitude correction.

3. Summary and discussion

Findings from this study so far are as follows:

Multi-model ensemble spread is considerably larger than the SEE-based (more likely realistic) spread when the models' differing biases are uncorrected. The ratio between the two spreads is about 1.5 to 1.8 before bias correction, and about 1.2 to 1.4 after individual model bias corrections.

The ratio of internal spread around individual model ensemble means (i.e., the spread of individual model ensemble members) to the standard error of estimate is in 0.8 – 1.0 range, showing slightly too tight an ensemble distribution. This result is expected in view of individual models having their own universe, and often (not always) recognizing less noise in that universe than there is in the real world.

Correcting forecasts so that the ratio of their interannual SD equals that of observations multiplied by their correlation skill (i.e., amplitude correction) makes less difference in the RMSE of the MME forecasts than model bias correction, but brings the spread of the MME forecasts within the neighborhood of that indicated by the skill-based SEE for intermediate and long leads. For shortest leads, the MME spread becomes smaller than the SEE-based spread.

A clear conclusion is that individual model biases should be corrected before the merging into a MME is done:

Correction of model amplitude biases should also be done. It reduces the interannual variability of the MME forecasts to be lower than that of the observations, to minimize squared errors and to create probabilistic reliability (lack of overconfidence). The lower the hindcast-based skill, the smaller the interannual variability of the MME forecasts should become.

A final thought concerns best way to display the forecast plume for users. Both the mean of the MME and the associated uncertainty must be shown in an easily understood and usable way.

References

Hayes, W. L., 1973: Statistics for the Social Scientists, Second Edition. Holt, Rinehart and Winston, Inc.

Kirtman, B. P., D. Min, J. M. Infanti, J. L. Kinter, D. A. Paolino, Q. Zhang, H. van den Dool, S. Saha, M. P. Mendez, E. Becker, P. Peng, P. Tripp, J. Huang, D. G. DeWitt, M. K. Tippett, A. G. Barnston, S. Li, A. Rosati, S. D. Schubert, Y.-K. Lim, Z. E. Li, J. Tribbia, K. Pegion, W. Merryfield, B. Denis and E. Wood, 2014: The US national multi-model ensemble for intra-seasonal to interannual prediction. *Bull. Amer. Meteor. Soc.*, **95**, in press.